A significance test in forward stepwise regression

Department of Statistics
Stanford University
Sequoia Hall
390 Serra Mall
Stanford, CA 94305, U.S.A.

Abstract: We extend the methods developed by Lockhart et al. (2013) and Taylor et al. (2013) on significance tests for penalized regression to stepwise model selection by iteratively applying the global null hypothesis test on the residual in a forward stepwise procedure. While not fully theoretically justified, the resulting method has the computational strengths of stepwise selection and partially solves the problem of inflated test statistics due to model selection. We illustrate the flexibility of this method by applying it to novel specialized applications of forward stepwise to a hierarchically constrained interactions model and generalized additive models.

AMS 2000 subject classifications: Primary 62M40; secondary 62H35.
Keywords and phrases: forward stepwise, model selection, significance test.

1. Introduction

Forward stepwise regression is a classical model selection procedure that begins with an empty model and sequentially adds the best predictor variable in each step. The usual χ^2 and F-tests for significance fail when a model has been selected this way and tend to be anti-conservative unless they are computed on a held-out validation dataset. In the lasso setting, Lockhart et al. (2013) derived a novel test statistic and its asymptotic null distribution, making possible valid inferences after model selection using the full data. Taylor et al. (2013) modified and extended those results to the group lasso (Ming and Lin, 2005) and other penalized regression problems, but only under the global null hypothesis. One of the strengths of these test statistics is that they can be used for valid significance testing when computed on the same data used for model selection, eliminating the need for data splitting. The present work iteratively applies the global null test of Taylor et al. (2013) for each step in forward stepwise selection, and works out some of the details necessary for models with grouped variables. The resulting method can be more statistically efficient than validation on held-out data, and more computationally efficient than penalized methods with regularization parameters chosen by cross-validation.

In Section 2 we establish notation and describe the forward stepwise algorithm used in most of the paper. Section 3 briefly reviews the parts of Lockhart et al. (2013) and Taylor et al. (2013) relevant to our significance test, and describes the group lasso which we require for applying the test with grouped

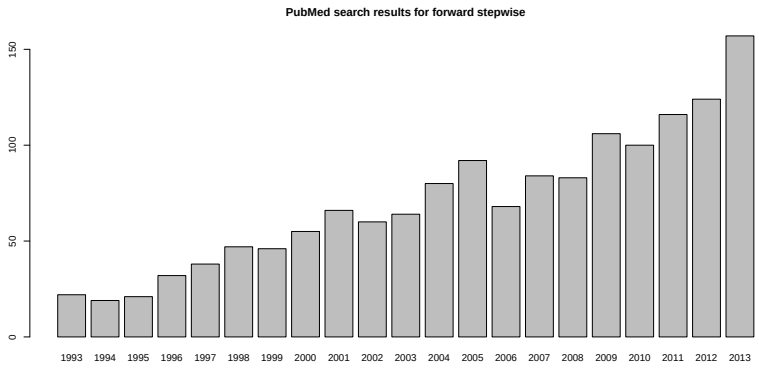


FIG 1. *Forward stepwise enjoys widespread use among practitioners*

variables. Simulation results in Section 4 show empirically that our method performs well in settings where forward stepwise itself performs well, and that various stopping rules using our test statistic–including some from Grazier G'Sell et al. (2013)–appear promising. In Section 5 we apply the method to several variants of forward stepwise tailored to models with interactions and generalized additive models, as well as to a real data example. (**To do:** Change this paragraph if necessary as the sections are completed)

2. Forward stepwise model selection

2.1. Background and notation

As a classical method dating back about half a century (see Hocking (1976) for a review), forward stepwise regression has not received much attention in recent years in the theoretical statistics community. But it continues to be widely used by practitioners. For example, search results on PubMed for forward stepwise (summarized in Figure 1) show that many recent papers mention the method and there is an increasing trend over time. Its popularity among researchers continues despite the fact that it invalidates inferences using the standard χ^2 or F-tests.

Some attempts to address this issue include Monte Carlo estimation of tables of adjusted F-statistic values (Wilkinson and Dallal, 1981), and permutation statistics (Forsythe et al., 1973). Aside from the works this paper is based on, there have been other recent attempts to do inference after model selection. Most of these make use of subsampling (Meinshausen and Bühlmann, 2010) or data splitting (Wasserman and Roeder, 2009). Our approach allows use of the full data and does not require the extra computation involved in subsampling. Before describing the full approach we first introduce notation and specify our

implementation of forward stepwise, which is slightly different from the most commonly used versions.

We allow forward stepwise selection to add groups of variables in each step, not only in the case of binary encoding for categorical variables but also for any grouping purpose. For example, groups of variables may be pre-designated factors such as expression measurements for all genes in a single functional pathway. To emphasize this we use g, h as indices rather than the usual i, j throughout. Since single variables can be considered groups of size 1, our general exposition includes non-grouped situations as a special case. Our variant of forward stepwise uses a different method than usual for choosing which group to add because we are using the same objective in the forward stepwise procedure that was used to derive the test statistics in penalized regression settings.

Denote $y \in \mathbb{R}^n$ for n i.i.d. measurements of the outcome variable. Let an integer $G \geq 2$ be the number of groups of explanatory variables. For each $1 \leq g \leq G$ the design matrix encoding the gth group is the $n \times p_g$ matrix denoted X_g, where p_g is the number of individual variables or columns in group g. When a group encodes a categorical variable as binary indicators for the levels of that variable, we use the full encoding with a column for every level. Although this introduces collinearity, our method does not require the design matrix to have full rank.

Define $p = \sum_{g=1}^{G} p_g$, the total number of individual variables, so $p = G$ in the case where all groups have size 1. Let X be the matrix constructed by column-binding the X_g, that is

$$X = \begin{pmatrix} X_1 & X_2 & \cdots & X_G \end{pmatrix}$$

We assume throughout that the individual columns of X are scaled to have unit 2-norm (we do not scale groups, this is accomplished later by weighting). With each group we associate the $p_g \times 1$ coefficient vector β_g, and write β for the $p \times 1$ vector constructed by stacking all of the β_g in order. Finally, our model for the response is

$$\begin{aligned} y &= X\beta + \sigma\epsilon \\ &= \sum_{g=1}^{G} X_g \beta_g + \sigma\epsilon \end{aligned} \tag{1}$$

where ϵ is noise. We assume Gaussian noise $\epsilon \sim N(0, \Sigma)$ with known covariance matrix Σ.

The model (1) is underdetermined when $p > n$. In such cases it still may be possible to estimate β well if it is sparse–that is, if it has few nonzero entries. In the rest of this paper we refer to variable groups X_g as noise groups if β_g is a zero vector and as true or signal groups if β_g has any nonzero entries. We refer to the number of such nonzero groups as the *sparsity* of the model, and denote this $k := \#\{g : \beta_g \neq 0\}$.

Before we describe the forward stepwise procedure we require one last ingredient. To each group of variables we assign a weight w_g. These weights act like penalties or costs, so increasing w_g makes it more difficult for the group X_g to

enter the model. The modeler can choose weights arbitrarily, but we mostly use one particular choice, based on p_g, that we discuss later. With this we are ready to describe the forward stepwise procedure.

2.2. Description of the algorithm

First, the user must specify the maximum number of steps allowed, which we denote *steps* (consistent with the R function for stepwise). To enable our test statistic computations, *steps* should be at most $\min(n, G) - 1$, but it is computationally desirable to set it as low as possible while still being larger than the (unknown) sparsity of β. This way forward stepwise has a chance of recovering all the nonzero coefficients of β and then terminating without performing much additional computation. In our implementation we treat the active set A as an ordered list to easily track the order of groups entering the model.

Algorithm 1 Forward stepwise variant with groups and weights

Input: An n vector y and $n \times p$ matrix X of G variable groups with weights w_g
Output: Ordered active set A of variable groups included in the model at each step

1: $A \leftarrow \emptyset$, $A^c \leftarrow \{1, \ldots, G\}$
2: **for** $s = 1$ to *steps* **do**
3: $\quad g^* \leftarrow \text{argmax}_{g \in A^c} \{\|X_g^T r_{s-1}\|_2 / w_g\}$
4: $\quad P_g \leftarrow I_{n \times n} - X_{g^*} X_{g^*}^{\dagger}$
5: $\quad A \leftarrow A \cup \{g^*\}$, $A^c \leftarrow A^c \backslash \{g^*\}$
6: $\quad$ **for all** $g \in A^c$ **do**
7: $\quad\quad X_g \leftarrow P_g X_g$
8: $\quad$ **end for**
9: $\quad r_s \leftarrow P_g r_{s-1}$
10: **end for**
11: **return** A

The active set A contains variable groups chosen to be included in the model. Fitting $\hat{\beta}$ can be done by tracking the individual fits and projections, or by simply fitting a linear model on the submatrix of X corresponding to A. We now use forward stepwise to refer specifically to Algorithm 1 unless otherwise specified. Note that other implementations of forward stepwise use different criteria for choosing the next variable to add, such as the correlation with the residual. Since we do not renormalize the columns after projecting the covariates (lines 6 to 8 above), and since we have weights, we are not computing correlations unless the design matrix is orthogonal and all weights are 1. There are advantages and disadvantages to both criteria which we do not discuss. Our choice was motivated by the group lasso result in Taylor et al. (2013), but we believe other criteria can be handled with appropriate modifications.

2.3. Performance of forward stepwise

Among model selection procedures, forward stepwise is one which performs variable selection: from a potentially large set of variables it chooses a subset to include in the model. The most general form of variable selection is called subset

selection, a procedure which searches among all 2^G subsets of G variable groups and picks the best model. This exhaustive search is computationally infeasible when G is large, and when possible it still runs the risk of over-fitting unless model complexity is appropriately penalized (as in (2) below). Forward stepwise produces a much smaller set of potential models, with cardinality at most *steps*. However it is a greedy algorithm, so the set of models it produces may not contain the best possible model. This is an inherent shortcoming of forward stepwise procedures and should be kept in mind when choosing between model selection methods.

So far we have left open the question of choosing among the models in the forward stepwise sequence, i.e. when to stop stepping forward. Some approaches for this problem can be posed as optimization criteria which stop at the step minimizing

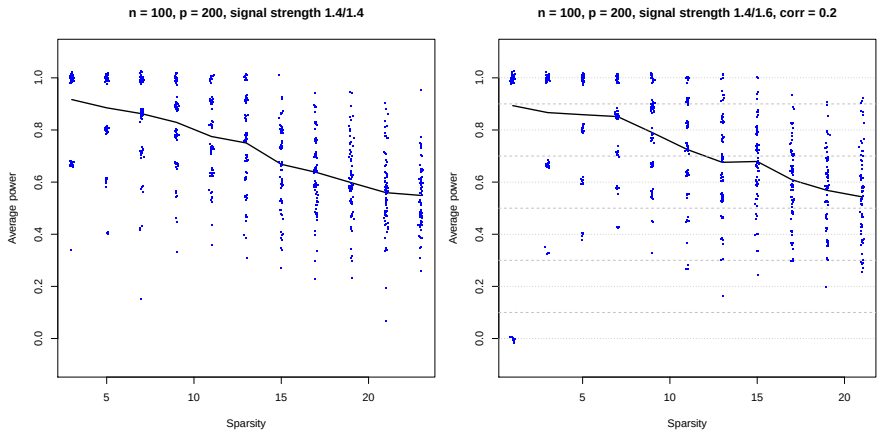
$$\frac{1}{2}\|y - X\beta_s\|_2^2 + \lambda\mathcal{P}(\beta_s) \tag{2}$$

(**To do:** describe construction of β after finding A) where we have written $\{\beta_s : s = 1, \ldots, steps\}$ as the sequence of models outputted by forward stepwise. The function $\mathcal{P}(\beta)$ is a penalty on model complexity usually taken to be the number of nonzero entries of β. Proposals for λ include 2 (C_p of Mallows (1973), AIC of Akaike (1974)), $\log(n)$ (BIC of Schwarz (1978)), and $2\log(p)$ (RIC of Foster and George (1994)). Stopping rules based on classical test statistics have also been used, so it is natural to consider using the new test statistics of Lockhart et al. (2013) or Taylor et al. (2013) to choose a model. Grazier G'Sell et al. (2013) examined some stopping rules using the asymptotic p-values of Lockhart et al. (2013) and showed their stopping rules control false discovery rate–the expected proportion of noise variables among variables declared significant (Benjamini and Hochberg, 1995). We explore this further in Section 4.

Although forward stepwise is a greedy algorithm producing a potentially sub-optimal sequence of models, under favorable conditions it can still perform well. There is a small segment of the compressed sensing literature (Donoho et al., 2006; Cai and Wang, 2011) with results stating that forward stepwise (often referred to in that literature as Orthogonal Matching Pursuit or OMP) can exactly select the true model under some stringent conditions involving quantities like the sparsity of the true model and the coherence of the design matrix. The coherence $\mu(X)$ of a matrix X with columns scaled to have unit 2-norm is defined as

$$\mu := \mu(X) = \max_{i \neq j}\{|\langle X_i, X_j\rangle|\} \tag{3}$$

Recalling that k is the sparsity of β, typical results in this literature say that if $k < (1/\mu + 1)/2$ and the nonzero coefficients of β are sufficiently large then forward stepwise recovers β with high probability. The coherence condition is necessary to guarantee exact recovery (Cai et al., 2010) in the sense that it is possible to construct counterexamples with $k = (1/\mu + 1)/2$. We refer the reader to the literature for details. For our purposes the conditions required to guarantee exact recovery are usually too stringent. Simulations show empirically



(a) Ternary design (b) Gaussian design with correlation

FIG 2. *The left panel shows the results of the simulation using design matrices with i.i.d. ternary entries taking values* $0, \pm 1$ *with equal probability. The left panel shows the results with design matrices of standard Gaussian entries with equi-correlation of 0.2.*

that forward stepwise can work well even when it is not working perfectly, and that it does so under a wide range of conditions.

For various sparsity levels k, Algorithm 1 was run on various data sets with *steps* equal to k and the "average power" was calculated as the number of true variables in the active set after k steps divided by k. (**To do:** Make this consistent with γ in simulations section) Nonzero coefficients had magnitude in a range of multiples of $\gamma := \sqrt{2\log(G)}$, e.g. in $[1.4\gamma, 1.6\gamma]$. Results are shown in Figure 2. After computing the coherence of these design matrices, some calculations show the required sparsity level to guarantee exact recovery in these situations is about 2 or smaller, and the required nonzero coefficient magnitude is likely in the range of 10 to 100 times γ. These simulations are far from the stringent conditions required by theory to guarantee perfect recovery, but the performance, while not perfect, may still be good enough for some purposes.

3. Significance testing: from lasso to group lasso

To do: Reorganize and update this section

In the ordinary least squares setting, a significance test for a single variable can be conducted by comparing the drop in residual sums of squares (RSS) to a χ^2_1 distribution. Similarly, when adding a group of k variables we can compare the drop in RSS to a χ^2_k random variable. This generally does not work when the group to be added has been chosen by a method that uses

the data, and in particular it fails for forward stepwise procedures adding the "best" (e.g. most highly correlated) predictor in each step. In that case, the test statistic (drop in RSS) does not match the theoretical null distribution even when the null hypothesis is true. Lockhart et al. (2013) introduced a new test statistic based on the knots in the lasso solution path. They derived a simple asymptotic null distribution, proved a type of convergence under broad "minimum growth" conditions, and demonstrated in simulations that the test statistic closely matches its asymptotic distribution even in finite samples. That work marked an important advance in the problem of combining inference with model selection. Taylor et al. (2013) extended that work to the group lasso (Ming and Lin, 2005) and other problems, and modified the test statistic to one with an exact finite sample distribution under the global null hypothesis.

Writing $\hat{\beta}(\lambda)$ for the lasso solution for a fixed value of λ, we need the following facts summarized in Lockhart et al. (2013) (ref Tibs2012?).

- The vector valued function $\hat{\beta}(\lambda)$ is a continuous function of λ. For the lasso path, the coordinates of $\hat{\beta}(\lambda)$ are piecewise linear with changes in slope occurring at a finite number of λ values referred to as *knots*. The knots depend on the data and are usually written in order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. We follow this convention.
- The *active set* A_k is a set of indices of variables for which the corresponding coordinates of $\hat{\beta}(\lambda_k)$ are potentially nonzero. Any variable with index not in A_k has a zero coefficient in $\hat{\beta}(\lambda_k)$, but the converse is not true. **To do:** is this equicorrelation set?
- Path algorithms for computing lasso solutions proceed by fitting models at a grid of λ values. The active set changes whenever λ crosses a knot, and predictor variables can both enter and leave the active set. However, at the first two knots λ_1 and λ_2 no variable can leave the active set. So the first two knots correspond to the first two variables entering the model.

Lockhart et al. (2013) prove that under the null hypothesis that A_k contains all the strong predictor variables, the distribution of a test statistic $T_k \propto \lambda_k(\lambda_k - \lambda_{k+1})$ is asymptotically Exp(1). In the lasso case we know a lot about the knots and active set, but the group lasso picture is slightly more complicated. For the group lasso, $\hat{\beta}(\lambda)$ does not have piecewise linear components. To overcome this difficulty we restrict our attention to the first group of variables to enter the active set since the analysis then follows almost exactly as for the lasso.

The *group lasso estimator* is a solution to the following convex problem

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|\beta_g\|_2 \tag{4}$$

The parameter $\lambda \geq 0$ enforces sparsity in groups: for large λ most of the β_g will be zero vectors. The weights w_g are usually taken to be $\sqrt{p_g}$ to normalize the penalty across groups. Note that this includes the usual lasso estimator as a special case when all of the groups are of size 1, since then the penalty term is the L^1-norm of β.

To do: Fix references below

The group lasso estimator is discussed in (ref that guy's thesis) and (ref Yuan and Lin). An important extension is the sparse group lasso (ref ???) which enforces sparsity in groups as well as sparsity of the coefficients within the groups. For a survey on group lasso and related factor models see (ref???). **To do:** review some more literature and add a few more references here if they seem worthwhile.

Before considering the group lasso, we review some ingredients of the proof for the lasso. Let $J = \{1, 2, \ldots, p\}$ index variables and consider a stochastic process $f_{j,s} = sX_j^T y$ defined on $T = J \times [-1, 1]$. This stochastic process is simply a collection of linear combinations of y, hence it is Gaussian under the assumption of Gaussian errors. The *Karush-Kuhn-Tuckher (KKT) conditions* (ref???) imply that $\lambda_1 = \max_j |X_j^T y|$. By introducing the sign variable s, we can remove the absolute value and write λ_1 as the maximum of our Gaussian process

$$\lambda_1 = \max_{(j,s)} f_{j,s} \tag{5}$$

We have exhibited the first knot as the maximum of a Gaussian process. We can do this for the second knot by introducing a new process. Let (j_1, s_1) be the maximizer so that $\lambda_1 = s_1 X_{j_1}^T y$, and define

$$\begin{aligned} f_{(j,s)}^{(j_1,s_1)} &= \frac{sX_j^T y - sX_j^T X_{j_1} X_{j_1}^T y}{1 - ss_1 X_j^T X_{j_1}} \\ &= \frac{sX_j^T (I - P_{j_1}) y}{1 - s_1 X_{j_1}^T X_j s} \end{aligned} \tag{6}$$

where P_j is the projection onto the subspace spanned by X_j. We can think of this as a "residual process" after regressing out the maximum. Write $M = \max_{j \neq j_1, s} f_{(j,s)}^{(j_1,s_1)}$, the maximum of this residual process. It can be shown from the KKT conditions that $M = \lambda_2$. To summarize, we have represented the first two knots of the lasso solution path as the maxima of some natural Gaussian processes. Distributional facts about Gaussian processes now allow us to make conclusions about the distribution of functions of the knots.

3.1. Group lasso

To extend this argument to the group lasso we need to define Gaussian processes that characterize the knots of the group lasso solution path.

To do: Either use the simplified argument in the case of equal weights (equal group sizes), or finish adjusting the proof of $M = \lambda_2$ to include the weights and use that version. The proof can go in an appendix.

To do: Modify write-up to match notation with Taylor et al. (2013) and then paste it in here

3.2. Relation to covariance test

Consider the simple case for the first covariate to enter the model. In this case the covariance test statistic is $T = \lambda_1(\lambda_1 - M)$. **To do:** Does $M = \lambda_2$? Check proof, now that statistic has changed

This is Lemma 5 from Lockhart et al. (2013).

- As in LTTT, $T = \lambda_1(\lambda_1 - M)$ and $M = \lambda_2$
- Convergence to the limiting Exp(1) distribution is too slow

$$\frac{P(\chi_k/w_g \geq m + t/m)}{P(\chi_k/w_g \geq m)} \to e^{-t} \text{ as } m \to \infty$$

  (when the group achieving λ_1 is group g and has rank k)
- The limiting distribution only depends on T, but we also observe M
- Let's just try the ratio (conditional χ_k tail probability) evaluated at T and M (it works better)
- Going one step further, instead of using the approximation (see LTTT Proof of Lemma 5)

$$\frac{M + \sqrt{M^2 + 4t}}{2} \approx M + \frac{t}{M}$$

  we can just use the left hand side
- For $T = \lambda_1(\lambda_1 - M)$ the left hand side simplifies to λ_1
- Now our p-value is

$$\frac{P(\chi_k/w_g \geq \lambda_1)}{P(\chi_k/w_g \geq \lambda_2)}$$

3.3. Exact p-value calculation

To do: Complete this section

We now describe how to calculate our p-value, following the discussion of group lasso in Taylor et al. (2013). For the rest of this section let $g = g^*$ be the index of the group attaining the maximum on line 3 of Algorithm 1. For a vector $u \in \mathbb{R}^p$, let u_h denote the coordinates of u corresponding to the columns of group h in the design matrix X. We can rearrange the columns of X to group these adjacently, so that

$$u^T X^T = \begin{pmatrix} u_1^T X_1^T & u_2^T X_2^T & \cdots & u_G^T X_G^T \end{pmatrix}$$

One step of the calculation will be to find an orthonormal basis for the linear space $\mathcal{L}_g = \{u \in \mathbb{R}^p : u_g^T X_g^T y = 0, u_h = 0 \text{ for all } h \neq g\}$ so that we can project orthogonally to this space. If X_g is a single column and $X_g^T y \neq 0$ (which should be the case since g maximizes the absolute value of this quantity), then the space is trivial and the desired orthogonal projection is the identity. Otherwise, if X_g has $p_g > 1$ columns the space $\mathcal{L}_g$ generally has dimension $p_g - 1$. We

compute an orthonormal basis by Gram-Schmidt and form a $p_g \times p_g$ matrix, which we denote V_g, by appending one column of zeroes (since Gram-Schmidt only produces $p_g - 1$ columns). With this we are ready to define the projection

$$P_g = \Sigma X_g V_g (V_g^T X_g^T \Sigma X_g V_g)^\dagger V_g^T X_g^T \tag{7}$$

Also define $H_g = (I - P_g)\Sigma$ and the conditional variance

$$\sigma_g^2 = y^T X_g X_g^T H_g X_g X_g^T y / (w_g^2 \|X_g^T y\|_2^2) \tag{8}$$

These simplify when X_g is a single column, in which case $P_g = 0$, $H_g = \Sigma$, and $\sigma_g^2 = X_g^T \Sigma X_g^T / w_g^2$. We now describe how to compute our p-value in Algorithm 2. Note that $F_{\chi_r^2}$ denotes the distribution function of a χ^2 random variable with r degrees of freedom.

Algorithm 2 Computing p-value

Input: Response y, grouped design matrix X with weights, inactive set A^c, index g of last group to enter active set A.
Output: Our p-value for the group g entering the model.

```
Compute H_g and σ_g^2
if X_g is a single column then
    σ_g^2 ← X_g^T Σ X_g / w_g^2
    X̃_g ← X_g / w_g · sign(X_g^T y)
else
    P_g ← Σ X V_g (V_g^T X^T Σ X V_g)^† V_g^T X^T
    H_g ← (I − P_g) Σ
    σ_g^2 ← y^T X_g X_g^T H_g X_g X_g^T y / (w_g^2 ‖X_g^T y‖_2^2)
    X̃_g ← H_g X_g X_g^T y / (‖X_g^T y‖_2 w_g)
end if
λ ← ‖X_g^T y‖_2 / w_g
a ← X^T (y − λ X̃_g)
b ← X^T X̃_g
Compute solution (v_−, v_+) of LinearFractional(a,b) optimization subproblem
r ← rank(X_g)
u ← [F_{χ_r^2}(v_−^2/σ_g^2) − F_{χ_r^2}(λ^2/σ_g^2)] / [F_{χ_r^2}(v_−^2/σ_g^2) − F_{χ_r^2}(v_+^2/σ_g^2)]
return u
```

We next describe the LinearFractional subproblem and its solution. We gave the problem this name because it originated in the form

$$\underset{h \neq g, \|u_h\|_2 = 1}{\text{maximize}} \; \frac{u_h^T X_h^T y - u_h^T X_h^T X_g X_g^T y}{1 - u_g^T X_g^T X_h u_h}$$

The solution we describe next is to a slightly different problem, and also incorporates the information that g maximizes $\|X_g^T y\|_2 / w_g$. For details see Appendix A or Taylor et al. (2013).

4. Simulations

To do: Compare to AIC and other classical stopping rules?

Algorithm 3 The LinearFractional optimization subproblem

Input: The p-vectors a, b in Algorithm 2, weights, inactive set A^c, a small tolerance number
Output: Solution pair (v_-, v_+).

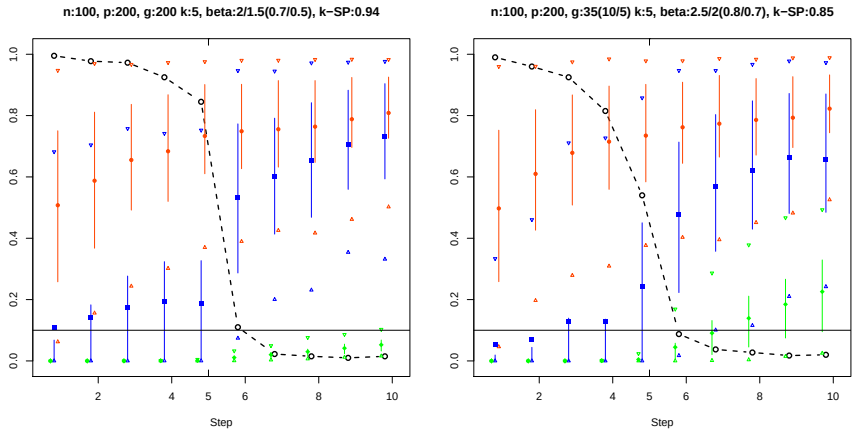
for h in A^c **do**
  if $\|b_h\|_2 == 0$ or $\|a_h\|_2/\|b_h\|_2 < \text{tol}$ **then**
    $(v_h^-, v_h^+) \leftarrow (0, \infty)$
  else
    $\theta_c \leftarrow a_h^T b_h/(\|a_h\|_2\|b_h\|_2)$
    $\theta_s \leftarrow \sqrt{1-\theta_c^2}$
    $\theta \leftarrow \arccos(\theta_c)$
    $\phi_s \leftarrow \theta_s\|b_h\|_2/w_h$
    if $\phi_s > 1$ **then**
      $(v_h^-, v_h^+) \leftarrow (0, \infty)$
    else
      $\phi \leftarrow \arcsin(\phi_s)$
      $z_\pm \leftarrow s_\pm\|a_h\|_2\cos(\phi)/(w_h - s_\pm\|b_h\|_2\cos(\theta-\phi))$ for $s_\pm = \pm 1$
      if $\|b_h\|_2 < w_h$ **then**
        $(v_h^-, v_h^+) \leftarrow (\max\{z_+, z_-\}, \infty)$
      else
        $(v_h^-, v_h^+) \leftarrow (\min\{z_+, z_-\}, \max\{z_+, z_-\})$
      end if
    end if
  end if
end for
$v^- \leftarrow \min_h v_h^-$
$v^+ \leftarrow \max_h v_h^+$
return (v^-, v^+)

We performed simulations with several classes of design matrices including Gaussian, categorical, and categorical from a real data set. Gaussian design matrices are generated either independently or with some global correlation $\rho > 0$ between all pairs of columns. Categorical matrices were generated by first choosing a vector of probabilities for a given variable from a Dirichlet prior, and then sampling category levels with that vector of probabilities. Resampling was used to ensure the minimum number of observations in any level was at least five. Finally, categorical varables were encoded as groups of 0-1 vectors using the full encoding. The categorical data matrices from real data are described in Section 5.3

Since we are interested in variable selection we generate signals with nonzero coefficients on the scale of $\gamma := \sqrt{2\log(p)/n}$. (**To do:** Check this scaling. **To do:** Cite some paper on support recovery?). Coefficients within a nonzero group have roughly the same magnitude, and the magnitudes for each nonzero group range from a lower limit times γ to an upper limit times γ. In plots, the limits are listed next to "beta:" and the numbers in parentheses are the largest and smallest 2-norms of nonzero coefficient groups respectively. Each plot also displays the number of observations or rows, n, the number of columns, p, the number of groups, g, and the largest and smallest group sizes in parentheses if the groups are not all size 1. The number of nonzero groups, k, is also displayed via inward-facing tick marks on the horizontal axes. The proportion of truly nonzero groups

recovered in the first k steps is denoted k-SP, where recall k is the number of truly nonzero groups.

Finally, most of the plots below show the following information. The horizontal axis is the step index for forward stepwise. The dashed line shows the proportion of iterations where a truly nonzero variable was added to the model at that step. Solid verticle lines show the middle 50% of p-values calculated at that step, with red calculated on a null model with no nonzero groups, blue calculated on the non-null model, and green calculated on the non-null model from the usual χ^2 significance test. The points in these lines are the average p-value at that step, and the triangles above and below show the 95% and 5% quantiles.



(a) Gaussian with correlations (b) Gaussian with groups

FIG 3. *The left panel shows results for Gaussian design matrices with equicorrelation $\rho = 0.2$ and equicorrelated noise with correlation 0.1. The right panel shows results for an independent Gaussian design with groups of sizes 5 and 10.*

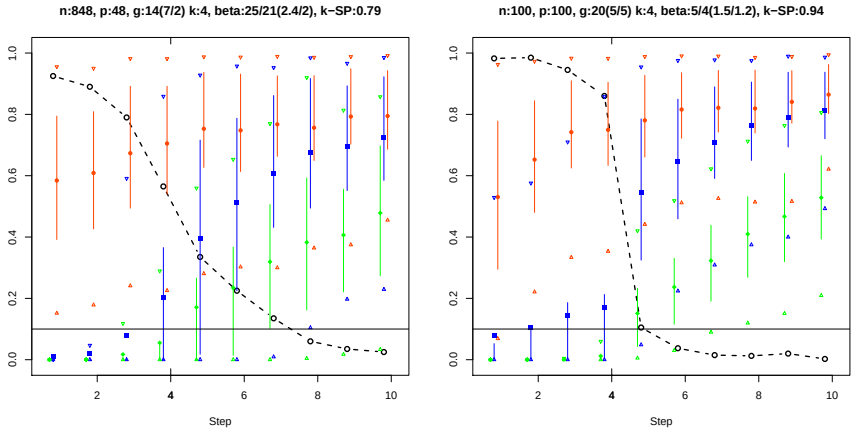
	Fdp	R	S	V	Power
(1) first	0.01	1.42	1.40	0.02	0.47
(1) forward	0.01	0.65	0.64	0.01	0.21
(1) last	0.04	2.97	2.78	0.19	0.93
(2) first	0.03	2.22	2.11	0.11	0.70
(2) forward	0.00	0.76	0.76	0.00	0.25
(2) last	0.06	3.22	2.98	0.24	0.99

TABLE 1

Evaluation of model selection using several stopping rules based on our p-values. The naive stopping rule performs well.

	RSS	Test Error	MSE(beta)
first	60.48	58.44	37.29
forward	60.66	58.27	53.57
last	0.93	2.40	11.16

TABLE 2
Prediction and estimation errors for a small simulation



(a) Real data categorical design (b) Simulated categorical design

FIG 4. *The left panel shows results for a categorical design taken from a real data set. The right panel shows results for categorical matrices with all categories having 5 levels.*

5. Applications

We now turn to applying forward stepwise and examining the behavior of our test statistic in several unique settings, including a real data example.

5.1. Glinternet for hierarchical interactions

In regression settings with many variables, considering models with pairwise interactions can drastically increase model complexity. Lim and Hastie (2013) propose a method called GLINTERNET to reduce the statistical complexity of this problem. The method imposes a strong hierarchical constraint on interactions (similar to that in Bien et al. (2012)) where an interaction term can only be included if both its main effects are also included. They accomplish this by creating a design matrix with both main effects alone and also with groups including main effects with their first order interactions. Then they fit a group lasso model with the expanded design matrix. Because interaction terms only appear in groups along with their respective main effects, the hierarchy condition holds for the fitted model. We now consider a related procedure as an example

	Fdp	R	S	V	Power
(1) first	0.07	1.00	0.80	0.20	0.13
(1) forward	0.00	0.60	0.60	0.00	0.10
(1) last	0.12	4.20	3.60	0.60	0.60
(2) first	0.00	2.20	2.20	0.00	0.37
(2) forward	0.00	0.80	0.80	0.00	0.13
(2) last	0.05	5.20	5.00	0.20	0.83

TABLE 3

Evaluation of model selection using several stopping rules based on our p-values. The naive stopping rule performs well.

problem, but first modify their method to simplify some parts. Let the expanded design matrix be given by

$$\tilde{X} = \begin{pmatrix} X_1 & \cdots & X_G & X_{1:2} & \cdots & X_{1:G} & X_{2:3} & \cdots & X_{(G-1):G} \end{pmatrix} \tag{9}$$

where $X_{g:h}$ is the submatrix including X_g, X_h, and all $p_g p_h$ column multiples between columns in group g and columns in group h. For example, if

$$X_g = \begin{pmatrix} X_{g1} & X_{g2} \end{pmatrix}, \quad X_h = \begin{pmatrix} X_{h1} & X_{h2} \end{pmatrix}$$

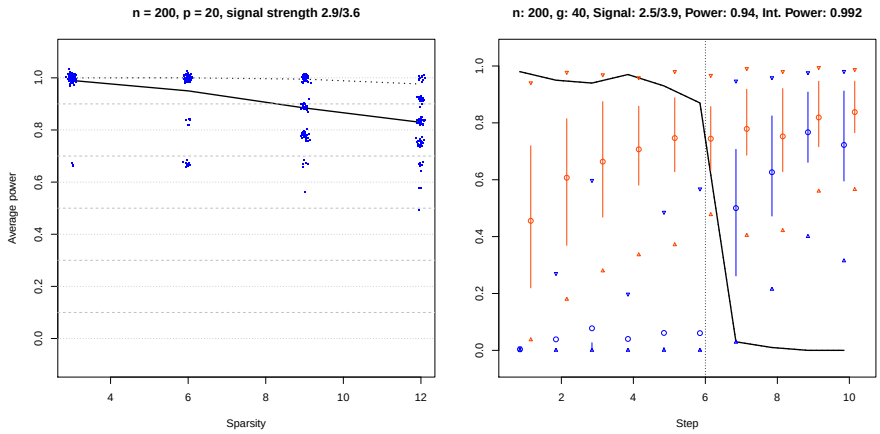
are two groups each containing two columns, then

$$X_{g:h} = \begin{pmatrix} X_g & X_h & X_{g1} * X_{h1} & X_{g1} * X_{h2} & X_{g1} * X_{h1} & X_{g2} * X_{h2} \end{pmatrix}$$

where * denotes the pointwise product (Hadamard product) of vectors (the ith entry of $X_{g1} * X_{h1}$ is the ith entry of X_{g1} times the ith entry of X_{h1}). Note that this expanded matrix has $\tilde{G} = G + \binom{G}{2} = O(G^2)$ groups. We refer to the first G of these as main effects or main effect groups, and the remaining as interaction groups. Finally, instead of fitting a model by group lasso, we use forward stepwise on the expanded design matrix. The overlapping groups still guarantee that our fitted model satisfies the strong hierarchy condition.

To demonstrate this method by simulation, we constructed signals which have the first $k/3$ main effects nonzero but with no interactions, and the remaining $2k/3$ nonzero main effects are matched to each other to form interactions. This way each nonzero main effect with any nonzero interactions has exactly one nonzero interaction. Furthermore, each nonzero interaction coefficient has been inflated to be larger than the corresponding main effect coefficients. This special case is favorable for our algorithm, but our purpose here is merely to demonstrate the flexibility of the hypothesis test and not to propose a general procedure for models with interactions.

Results are shown in Figure 5. The left panel shows average power of forward stepwise. Power is calculated using the group structure we impose, and not in terms of the original main effects and interactions. However, the dotted line shows a more forgiving definition of power where we are rewarded for discovering part of a nonzero group, i.e. for discovering only one main effect from a true interaction group. In the right panel we see that the global null result still holds (red vertical lines), and that with a signal the p-value is small until the signal has been recovered (blue). The interaction power is the proportion of nonzero interaction groups that were discovered and is labeled "Int. Power."



(a) Forward stepwise on glinternet (b) P-value with glinternet

FIG 5. *The left panel shows the power of forward stepwise for various sparsity levels. The right panel shows our p-value marginally over each step, with red indicating the global null case and blue indicating a signal with* $k = 6$ *nonzero groups.*

5.2. Generalized additive model selection

To do: Finish code
To do: Write section

5.3. Real data example

To do: Sanity check my analysis
To do: For rare mutations try binary wildtype/non-wt?
To do: Finish writing section

Rhee et al. (2006) use genomic information to predict efficacy of antiretroviral drugs in treating HIV. Quantitative measurements of drug response/resistance were regressed on categorical covariates encoded the presence and type of certain genetic mutations. We attempt a similar analysis using forward stepwise, and report our P-value at each step. Categorical covariates are encoded as groups of binary variables using the full encoding, and weights are set to the square-root of the number of levels for that covariate. We perform forward stepwise once for each drug response, restricting to the subset of the data with no missing values and requiring all categorical variables to have 2 or greater observations in each level.

Results from two data sets are displayed in Figure 6 and Figure 7. The PI data contains protease inhibitor resistance mutations, and the NRTI data

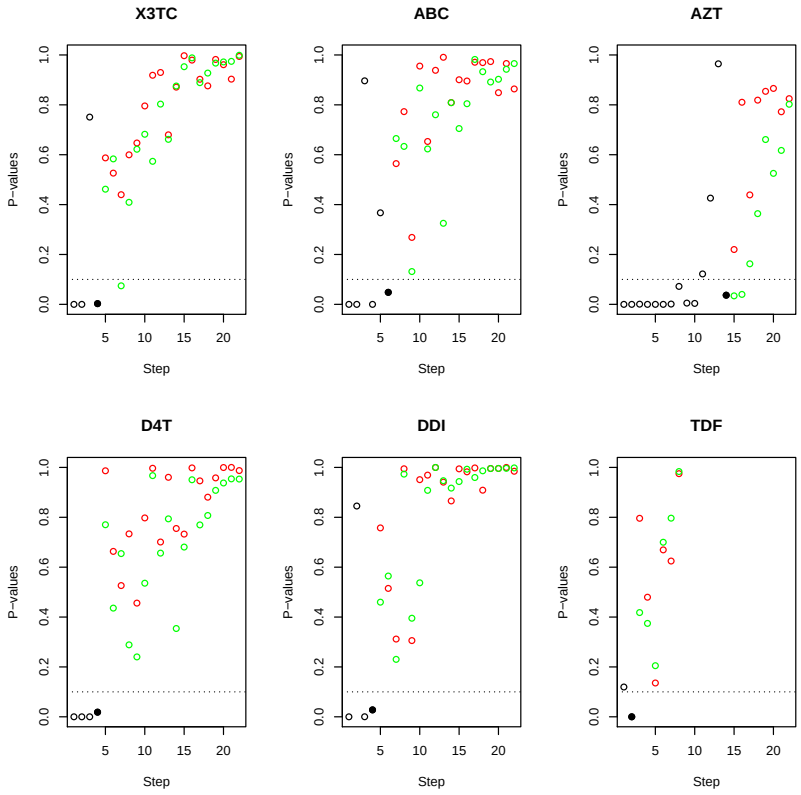


FIG 6. *Results from PI dataset*

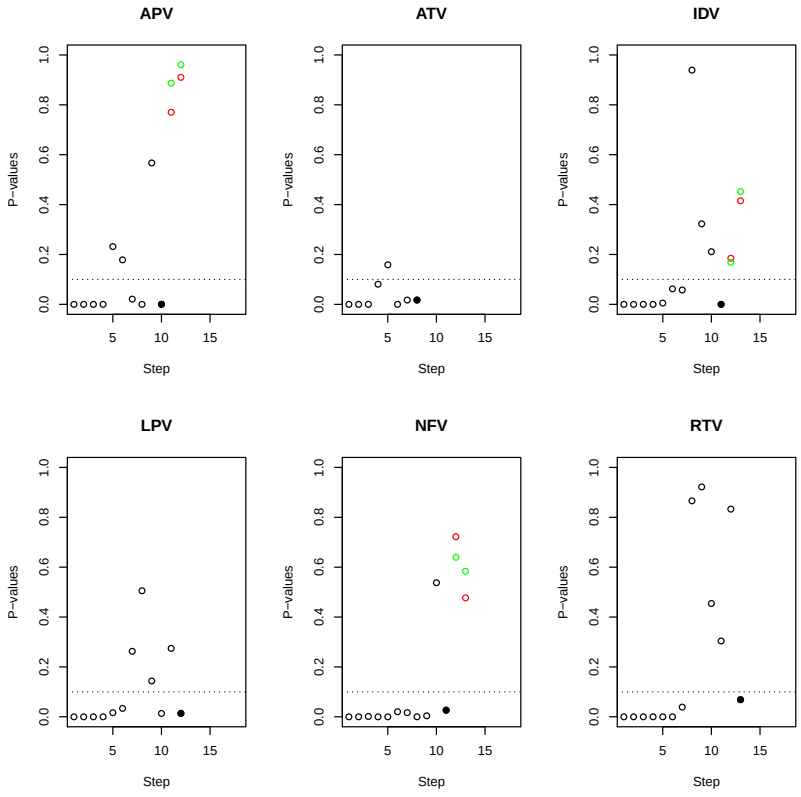


FIG 7. *Results from NRTI dataset*

contains nucleoside RT inhibitor resistance mutations. Each panel shows results for a different drug, with p-values plotted by step in forward stepwise. The solid black point is the last p-value to fall under the 0.1 threshold and points after it are colored red. If the *last* stopping rule were used then the p-values for including variables are the black points, and the resulting models are shown in Tables 4 and 5. (**To do:** Fix this ref after re-running)

	Drug	n	G	Selected variables
1	X3TC	633	24	P36 P184 P181 P151
2	ABC	628	24	P128 P184 P74 P151 P44 P218
3	AZT	630	24	P128 P44 P184 P218 P151 P60 P77 P111 P202 P181 P82 P74 P225 P50
4	D4T	630	24	P128 P151 P44 P218
5	DDI	632	24	P36 P151 P184 P74
6	TDF	353	9	P225 P184

TABLE 4
Variables chosen using last stopping rule

	Drug	n	G	Selected variables
1	APV	768	13	P6 P90 P46 P54 P76 P24 P32 P50 P73 P11
2	ATV	329	9	P30 P90 P54 P46 P24 P50 P32 P73
3	IDV	827	14	P6 P90 P46 P54 P73 P50 P24 P62 P32 P88 P76
4	LPV	517	13	P6 P90 P54 P46 P76 P32 P62 P50 P73 P11 P88 P30
5	NFV	844	14	P6 P90 P46 P30 P54 P88 P24 P73 P50 P62 P32
6	RTV	795	14	P6 P90 P54 P46 P24 P32 P50 P62 P73 P30 P11 P88 P76
7	SQV	826	14	P6 P90 P54 P24 P73 P88 P46 P30 P11 P50 P76

TABLE 5
Variables chosen using last stopping rule

6. Discussion

To do: Brief summary
To do: After finising everything else, move some points of discussion here
To do: Mention ongoing work, like tracking all the constraints to get an exact p-value at every step.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19** 716–723.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Bien, J., Taylor, J. and Tibshirani, R. (2012). A lasso for hierarchical interactions. *arXiv:1205.5050*. Submitted to Annals of Statistics., URL http://arxiv.org/abs/1205.5050.

Cai, T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *Information Theory, IEEE Transactions on*, **57** 4680–4688.

Cai, T. T., Wang, L. and Xu, G. (2010). Stable recovery of sparse signals and an oracle inequality. *IEEE Trans. Inf. Theor.*, **56** 3516–3522. URL http://dx.doi.org/10.1109/TIT.2010.2048506.

Donoho, D., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, **52** 6–18.

Forsythe, A. B., Engelman, L., Jennrich, R. and May, P. R. A. (1973). A stopping rule for variable selection in multiple regression. *Journal of the American Statistical Association*, **68** 75–77. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1973.10481336.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, **22** pp. 1947–1975. URL http://www.jstor.org/stable/2242493.

Grazier G'Sell, M., Wager, S., Chouldechova, A. and Tibshirani, R. (2013). False Discovery Rate Control for Sequential Selection Procedures, with Application to the Lasso. *ArXiv e-prints*. 1309.5352.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, **32** pp. 1–49. URL http://www.jstor.org/stable/2529336.

Lim, M. and Hastie, T. (2013). Learning interactions through hierarchical group-lasso regularization. *ArXiv e-prints*. 1308.2719.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2013). A significance test for the lasso. *arXiv:1301.7161*. Submitted to Annals of Statistics, URL http://arxiv.org/abs/1301.7161.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, **15** pp. 661–675. URL http://www.jstor.org/stable/1267380.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72** 417–473.

Ming, Y. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68** 49–67.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L. and Shafer, R. W. (2006). Genotypic predictors

of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, **103** 17355–17360. http://www.pnas.org/content/103/46/17355.full.pdf+html, URL http://www.pnas.org/content/103/46/17355.abstract.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6** pp. 461–464. URL http://www.jstor.org/stable/2958889.

TAYLOR, J., LOFTUS, J. and TIBSHIRANI, R. (2013). Tests in adaptive regression via the kac-rice formula. *ArXiv e-prints*. 1308.3020.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, **37** 2178–2201. Zentralblatt MATH identifier: 05596898; Mathematical Reviews number (MathSciNet): MR2543689, URL http://projecteuclid.org/euclid.aos/1247663752.

WILKINSON, L. and DALLAL, G. E. (1981). Tests of significance in forward selection regression with an f-to-enter stopping rule. *Technometrics*, **23** pp. 377–380. URL http://www.jstor.org/stable/1268227.

Appendix A: Derivation of closed forms of statistic

To do: Clean this section
To do: Make $M = \lambda_2$ a special case corollary

For the sake of completeness this appendix provides full derivations of closed forms for the quantities required to compute our statistic. We require two facts from Taylor et al. (2013)

A point $\eta \in \mathcal{K}$ maximizes f_η over a convex set $\mathcal{K}$ if and only if the following conditions hold:

$$\nabla f_{|T_\eta \mathcal{K}} = 0, \quad \widetilde{f}_\eta^\eta \geq \mathcal{V}_\eta^+, \quad \widetilde{f}_\eta^\eta \leq \mathcal{V}_\eta^-, \quad \text{and} \quad \mathcal{V}_\eta^0 \leq 0. \tag{10}$$

The same equivalence holds true even when $\mathcal{K}$ is only locally convex.

Write $M_{\eta,h}^{\pm}$ and $M_{\eta,h}^0$ as the corresponding suprema and infima over the restricted parameter set S_h. Note that the characterization of the global maximizer (Lemma 1, general paper) implies $M_{\eta,h}^0 \leq 0$ and $M_{\eta,h}^+ \leq M_{\eta,h}^-$ for all $h \neq g$. This will be used below to eliminate some degenerate cases of the optimization sub-problem on each S_h.

Now fix h, let $P_g = X_g X_g^T$ and define

$$\begin{aligned} a_h &= w_h^{-1} X_h^T (I - P_g) y \\ b_h &= w_g w_h^{-1} X_h^T P_g y / \|X_g^T y\| \\ c_h &= a_h^T b_h / (\|a_h\| \|b_h\|) \\ K_h &= \{x : \|x\| = 1, b_h^T x > w\} \end{aligned}$$

Note that $K_h = \emptyset$ if and only if $\|b_h\| < w$. We consider two cases. First, suppose $\|b_h\| > w$. In this case we must rule out one sub-case, when a_h is not in the polar cone of K_h. If this occurs, then on at least one side of the boundary of K_h we have $a_h^T x > 0$ there, so that the infimum inside K is $-\infty$ but the supremum outside K is $+\infty$. This violates the characterization of the global maximizer of the original process as described above, so this case cannot occur.

Now, if a_h is in the polar cone of K_h (and ruling out the probability zero case where $a_h^T x = 0$ on the boundary of K_h), then the numerator $a_h^T x < 0$ on K_h so there is a finite positive infimum in the interior K_h. Similarly there is a finite positive supremum on the interior of K_h^c (the numerator is negative near the boundary of K_h by continuity).

To attain the infimum on K_h requires making x close to b_h and simultaneously making the numerator closer to zero, so x should be on the side of b_h that is closer to a_h. To attain the supremum on K_h^c requires making x simultaneously close to a_h (so that $a_h^T x > 0$) and b_h (to make the denominator small). In summary, both the infimum and supremum are attained with x between a_h and b_h, when the angle θ between a_h and b_h is larger than both the angle ψ between a_h and x and the angle ϕ between x and b_h. This allows the simplification $\phi = \theta - \psi$. This is easier to verify for the case when $\|b_h\| < w$.

We have verified that solving both cases requires finding the maxima of the trigonometric form $\|a_h\|\cos(\psi)/(w - \|b_h\|\cos(\theta - \psi))$ of the linear fraction on the interiors of K_h (if it is nonempty) and K_h^c. The derivative is proportional to

$$\frac{\|b_h\|\sin(\theta) - w\sin(\psi)}{w - \|b_h\|\cos(\theta - \psi)}$$

Since we have ruled out the boundary of K_h, we can ignore the denominator and focus on critical points corresponding to the angles $\psi^{\pm}$ (symmetric about $\pi/2$) where $\sin(\psi^{\pm}) = (\|b_h\|/w)\sin(\theta)$. That is,

$$\psi^{\pm} \in \arcsin \frac{\|b_h\|}{w}\sqrt{1 - c_h^2}$$

Note that these critical points only exist if $\sin(\theta) < w/\|b_h\|$, and since we have argued that the infimum and supremum are attained it follows that this condition is met in our case. Let ψ^+ denote the smaller of the two angles. If K_h is nonempty, then ψ^- gives the infimum in K_h and ψ^+ the supremum on K_h^c. This is necessary since if x is in K_h then its angle from a_h is larger than $\pi/2$ because a_h is in the polar cone of K_h. If K_h is empty, then the supremum is still attained at ψ^+ since the numerator is negative at $\psi^- > \pi/2$ and the denominator is always positive.

Finally we calculate the values of the linear fraction at $\psi^{\pm}$. Let us first record some facts:

$$\begin{aligned}
0 < \psi^+ < \pi/2, \quad \psi^- &= \pi - \psi^+ \\
\cos(\theta) = c_h, \quad \sin(\theta) &= \sqrt{1 - c_h^2} \\
\sin(\psi^+) = \sin(\psi^-) &= \frac{\|b_h\|}{w}\sin(\theta) \\
\cos(\psi^+) = -\cos(\psi^-) &= \sqrt{1 - (\|b_h\|/w)^2(1 - c_h^2)}
\end{aligned}$$

Using the angle difference formula,

$$\begin{aligned}
\cos(\theta - \psi^+) &= \frac{\|b_h\|}{w}(1 - c_h^2) + c_h\cos(\psi^+) \\
&= \frac{\|b_h\|}{w}(1 - c_h^2) + c_h\cos(\psi^+)
\end{aligned}$$

Hence

$$M_{\eta,h}^+ = \frac{\|a_h\|c_h}{w - (\|b_h\|^2/w)(1 - c_h^2) - \|b_h\|c_h\cos(\psi^+)}$$

And $M_{\eta,h}^-$ is ∞ when $\|b_h\| < w$ and otherwise given by

$$M_{\eta,h}^+ = \frac{\|a_h\|c_h}{w - (\|b_h\|^2/w)(1 - c_h^2) + \|b_h\|c_h\cos(\psi^+)}$$

I haven't been able to algebraically simplify these yet to a form that allows my λ_2 proof below to go through. The version below just assumes $\|b_h\| < w = 1$.

We can rewrite this by rationalizing the denominator:

$$M = \frac{a_h^T b_h + \sqrt{a_h^T a_h (1 - b_h^T b_h) + (a_h^T b_h)^2}}{1 - b_h^T b_h}$$

Leaving M in this form we now consider λ_2. The KKT conditions for the group lasso problem give (directly from Yuan and Lin)

$$\|X_g^T(y - X\beta)\| \leq \lambda w_g \quad \forall \beta_g \neq 0$$
$$\beta_g = \left(1 - \frac{\lambda w_g}{\|S_g\|}\right)_+ S_g$$

where $S_g = X_g^T(y - X\beta_{-g})$. These expressions simplify for $\lambda_2 \leq \lambda < \lambda_1$. Letting g be the index of the first group to enter we now have $S_g = X_g^T y$

$$\beta_g = \left(1 - \frac{\lambda w_g}{\|X_g^T y\|}\right) X_g^T y$$
$$\lambda w_h \geq \|X_h^T y - X_h^T P_g y (1 - \lambda w_g / \|X_g^T y\|)\| \quad \forall h \neq g$$

Now let $\lambda = \lambda_2$, so the inequality above is strict for all other groups and equality is attained for the second group to enter. Hence λ_2 satisfies

$$\begin{aligned}\lambda_2 &= \max_{h \neq g} \|X_h^T (I - P_g) y - \lambda_2 (w_g X_h^T P_g y / \|X_g^T y\|)\| / w_h \\ &= \max_{h \neq g} \|a_h - \lambda_2 b_h\|\end{aligned}$$

Write $\lambda_{2,h}$ as the positive root of the quadratic equation obtained by squaring the norm, so $\lambda_{2,h}^2 = \|a_h - \lambda_2 b_h\|^2$. Then $\lambda_2 = \max_{h \neq g} \lambda_{2,h}$. Solving this (note that $b_h^T b_h < 1$), we find

$$\lambda_{2,h} = \frac{a_h^T b_h + \sqrt{a_h^T a_h (1 - b_h^T b_h) + (a_h^T b_h)^2}}{1 - b_h^T b_h}$$

Hence when h is the index of the second group to enter, we have $\lambda_2 = \lambda_{2,h} = M$.